ChinaXiv合作期刊

科技文献语篇元素自动标注模型研究综述*

于改红[1,2]　张智雄[1,2,3]　马娜[1,2]

[1] 中国科学院大学　北京 100049　[2] 中国科学院文献情报中心　北京 100190

[3] 中国科学院武汉文献情报中心　武汉 430071

摘要：[目的/意义]为更好地提升科技文献的语义丰富化效果，对国内外科技文献语篇元素标注模型、技术和方法进行调研总结，为文本挖掘、科技论文知识抽取、语义分析系统研究者提供借鉴。[方法/过程]利用学术网站搜索和相关数据库搜索引擎，对涉及科技论文标注、语篇元素、知识抽取、句子识别和自动文章分类等参考文献以及研究报告进行深入阅读和调研，对语篇元素自动标注模型以及相关工作进展进行研究总结。[结果/结论]科技文献语篇元素标注具有非常重要的实际应用价值，构建标注模型需充分考虑构建思想、标注领域和标注粒度以及标注技术手段等方面。

关键词：科技文献　语篇元素　标注模型　自动标注

分类号：G251

DOI：10.13266/j.issn.0252-3116.2018.15.015

chinaXiv:202308.00603v1

1 引言

过去几年中，科技文献的激增促进了文本挖掘工具的兴起，使得快捷地从文本中抽取有用的知识工具成为可能。如何帮助用户快速找到其想要的文献、如何找到文献中特定类别信息（比如所有实验数据）、如何获取特定类别的知识单元显得更有意义和前瞻性，如何快速并且全面地获取用户想要的研究信息越来越成为亟需解决的问题，并且是非常有意义的工作。

本研究将语篇元素定义为能够明确表示对科技文献中蕴含的知识价值进行功能描述的片段，其可以是一个从句、一个完整句子、一个段落，甚至一个片段，本文对语篇元素的标注信息定义为对其蕴含的语义类别信息进行标注，如研究思路、理论工具和方法、科学试验、实验结果、研究结论等。如何将论文中上述有价值的语义知识揭示出来，让其能够被方便地发现和使用，已经成为当前数字图书馆研究的一个重要课题。近年来，来自数字图书馆、知识抽取、知识组织和揭示等领域的专家学者从不同角度开展研究，但大多局限于对科技论文中的零散知识点及其关系进行标注和揭示，如实体抽取和关系抽取等，从整体上有效揭示科技论文中隐藏的丰富语义知识内容较为欠缺，要做好这项工作的基础就是要对科技文献语篇元素建立有效的标注模型，对上述语篇元素的知识进行组织揭示。自动标注模型是进行语篇元素自动标注工作的数据处理基准和数据组织规范框架，所有的标注工作都是建立在模型之上。因此，本文调研了国际知名研究学者和重要实验室在科技文献语篇元素结构自动标注的工作和研究进展，并着重对科技文献标注模型进行分析总结，作为重要的任务研究实施。

本文基于学术网站搜索和数据库搜索引擎对“discourse annotation scheme, automatic annotation, semantic annotation, sentence classification, 语篇结构，句子分类，自动分类，语义标注”等关键术语进行检索以及关联阅读，然后详细分析研究了近 15 年的 50 多篇研究文献，分别来自英国、美国、中国等国家和欧盟，最后整理出专门针对语篇元素自动标注模型的几个有影响力的研究团队作为本文重点综述分析的对象。下文将首先对典型的语篇元素标注模型进行详细描述，包括对瑞士 H. Ribaupierre 等[1-6]、英国华威大学 M. Liakata 等[7-19]、牛津大学 S. Teufel 等[20-25]、普拉大学 F.

* 本文系中国科学院文献情报能力建设专项项目“基于 arXiv 数据的物理领域科研论文自动语义标注和索引应用示范”（项目编号：院 1657）研究成果之一。

作者简介：于改红（ORCID:0000-0003-1301-2871），馆员，硕士；张智雄（ORCID:0000-0003-1596-7487），中国科学院武汉文献情报中心主任，研究员，博士，通讯作者，E-mail:zhangzhx@mail.las.ac.cn；马娜（ORCID:0000-0001-5016-0879），馆员，硕士。

收稿日期：2017-12-20　修回日期：2018-03-16　本文起止页码：132-144　本文责任编辑：刘远颖

Ronzano 等[26-31]相关研究人员及其团队的研究模型进行总结介绍，主要包括对概念层、元数据层、文章结构层、话语修辞层、关系层次的详细说明；其次深入分析比较各个模型的不同，总结出面向科技文献语义标注的多层次语义标注模型需要考虑的角度和方面，如构建思想、任务驱动、标注粒度、研究领域等，以帮助科研人员更好地建立和选择模型；最后对本文工作进行总结，并对接下来的工作进行展望。

2 典型的科技文献语篇元素标注模型描述

2.1 SciAnnoDoc 模型

SciAnnoDoc 模型是由来自牛津大学的 H. Ribaupierre 等研究提出。该研究的任务是提高信息检索的精准率和提升科技文献搜索引擎的使用效果[1-2]，该研究假设当科学家在检索信息的时候，他们通常都有一个精确的检索目标，用户不是去检索"关于主题 T"的文献，而是试图去回答特定的问题，比如找到一个概念的定义，寻找特定问题的结果，检查一个思路是否被证实，或者比较两篇文章的科学结论。回答这些精确或复杂的科学论文的查询需要对文章的全部内容进行精确建模标注，尤其是对每一篇文章的语篇类型进行标注。

该团队经过反复多次对科学家进行问卷调查和专家论证判断，提出了以用户为中心的 SciAnnoDoc 科技文献标注模型[3-5]，对语篇元素进行建模。该模型将科技全文分为 4 个层次进行标注，包括概念层、元数据层、修辞话语层、引用关系层。如图 1 所示[5]：

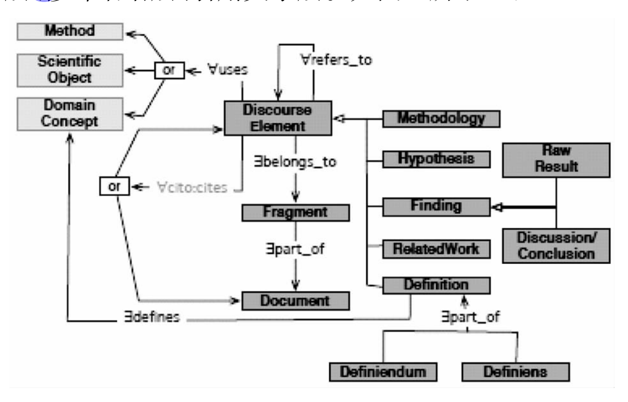


图 1 SciAnnoDoc 模型

(1)概念层(Domain Concept)：对文章本体或者描述科技术语词表或者文章中的概念进行标注。

(2)元数据层(Metadata)：描述元数据文本信息，比如作者、出版年、发表期刊或会议信息等。

(3)修辞话语层(Discourse Element)：这是每个模型的重点组成部分，描述元素发挥的作用和包含的知识内容属性，分解为 5 个方面：发现(Findings)、假设(Hypothesis)、方法(Methodology)、相关工作(Related work)以及定义(Definition)。

(4)引用关系层(relation)：描述文章之间的引用和关联关系。

2.2 CoreSC 模型

CoreSC 模型的发展成型主要经历了两个关键阶段的研究：第一阶段的科技文献核心信息 CISP 元数据模型(core information about scientific papers)[8]和第二阶段的核心科技概念模型 CoreSC(core scientific concepts)[11]。

第一阶段 CISP 元数据模型来源于 EXPO[7]中描述通用科技概念的子类，其主要包含了描述一项科学调查研究至关重要的概念，对概念类别经过专家调研和实际论文标注分析，精炼为以下 12 个类别最终作为 CISP 的模型分类：研究目标(goal of investigation)、研究对象(object of investigation)、研究方法(method of investigation)、实验(experiment)、观察(observation)、假设(hypothesis)、结果(results)、结论(conclusion)、动机(motivation)、背景(background)、问题(problem)、例子(example)，其中 8 个核心类为研究目标、研究动机、研究对象、研究方法、实验、结果、观察和结论[8-9]。

第二阶段，核心科技概念 CoreSC 模型，是在 CISP 基础上丰富完善的，于 2010 年被正式提出，旨在自动识别发现文章中一项科研调查的组成部分，是句子级的文本标注模型，具体模型描述如表 1 所示[11-12]，主要包含了 3 个层次的标注，表 1 展示了第一个层次的 11 个类别的含义和第二个层次的类别属性(New、Old、Advantage、Disadvantage)。

(1)修辞类别层：第一个层次包含了 11 个修辞类别，包括假设(Hypothesis)、动机(Motivation)、背景(Background)、目标(Goal)、对象(Object)、方法(Method)、实验(Experiment)、模型(Model)、观察(Observation)、结果(Result)、结论(Conclusion)。

(2)概念类别属性层：第二个层次是对概念属性的标注，如 New 或 Old 标注了一项方法是新方法还是旧方法，Advantage 或 Disadvantage 标注了一项方法的优势和劣势。

(3)概念识别 ID 层：第三个层次是 ConceptID 标识相同概念的相关联实例集合，如所有属于相同方法的句子关联在一起使用相同的 ConceptID。

chinaXiv:201903.00063v1

ChinaXiv合作期刊

表 1　核心科技概念(CoreSC)标注模型

类别	功能描述
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	the data/phenomena recorded in an investigation
Result	factual statements about the outputs of an investigation
Conclusion	statements inferred from observations & results relating to research hypothesis

2.3　Argumentative Zoning——AZ 模型

S. Teufel 的论证分区 AZ 模型[20,23]是受到知识声明概念的启发"写文献的行为与声明一条新知识的所有权相关,是经过作者领域的同行评审之后加入科技知识库和出版的行为",其中心思想是假设科技文献包含了对其他贡献者的积极和消极的陈述,因此模型更关注于知识声明(Knowledge claim)的组织揭示。

AZ 模型的发展同样经历了两个关键的阶段,最初1999 年 S. Teufel 等把文献分成 7 个分区,具体模型描述见图 2[20],OTHER、OWN 和 BACKGROUND 分别关联于这些片段的知识所有权归属,BASIS 声明了使用其他工作作为当前工作研究基础或出发点或获得的支持,CONTRAST 包含了对不同研究工作之间的比较(比如指出其他工作的不足),AIM 指出了文章的主要知识声明,TEXTUAL 给出了文本的物理位置信息。

BASIC SCHEME	BACKGROUND	Sentences describing some (generally accepted) background knowledge	FULL SCHEME
	OTHER	Sentences describing aspects of some specific other research in a neutral way (excluding contrastive or BASIS statements)	
	OWN	Sentences describing any aspect of the own work presented in this paper – except what is covered by AIM or TEXTUAL, e.g. details of solution (methodology), limitations, and further work.	
	AIM	Sentences best portraying the particular (main) research goal of the article	
	TEXTUAL	Explicit statements about the textual section structure of the paper	
	CONTRAST	Sentences contrasting own work to other work; sentences pointing out weaknesses in other research; sentences stating that the research task of the current paper has never been done before; direct comparisons	
	BASIS	Statements that the own work uses some other work as its basis or starting point, or gets support from this other work	

Figure 1: Overview of the annotation scheme

图 2　论证分区 AZ 模型

chinaXiv:202308.00603v1

于改红，张智雄，马娜. 科技文献语篇元素自动标注模型研究综述[J]. 图书情报工作,2018,62(15):132－144.

AZ 模型一直不断被学术界丰富完善并使用，直到 2009 年发展到 AZ-II 模型[23]，扩展到了 15 个类别，对比于原始 AZ 模型，AZ-II 扩展模型的变化在于：

(1)类别 AIM 保持一致；

(2)类别 BACKGROUND 被重新命名为 CO_GRO 或者成为通用背景；

(3)类别 OTHER 被细分为其他人的工作(OTHER)和作者自己之前的工作(PREV_OWN)；

(4)类别 BASIS 被细分为使用(USE)和支持(SUPPORT)；

(5)类别 CONTRAST 被细分为中立对比(CODI)、矛盾对立(ANTISUPP)、结合研究不足评论(GAP_WEAK)；

(6)类别 OWN 被细分为方法描述(OWN_MTHD)、结果(OWN_RES)、结论(OWN_CONC)以及作者指出可修复的错误信息(OWN_FAIL)；

(7)停止使用类别 TEXTUAL，因为对比其他类别该类别信息量更少。

该模型引入了两个新的类别——新知识声明的优势(NOV_ADV)和未来工作限制声明(FUT)。具体类别含义如表 2 所示[22－23]：

表 2　论证分区 AZ-II 标注模型

类别	功能描述
AIM	Statement of specific research goal, or hypothesis of current paper
NOV_ADV	Novelty or advantage of own approach
CO_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper)
OTHR	Knowledge claim (significant for paper) held by somebody else. Neutral description
PREV_OWN	Knowledge claim (significant) held by authors in a previous paper. Neutral description.
OWN_MTHD	New Knowledge claim, own work: Methods
OWN_FAIL	A solution/method/experiment in the paper that did not work
OWN_RES	Measurable/objective outcome of own Work
OWN_CONC	Findings, conclusions (non-measurable) of own work
CODI	Comparison, contrast, difference to other solution (neutral)
GAP_WEAK	Lack of solution in field, problem with other solutions
ANTISUPP	Clash with somebody else's results or theory; superiority of own work
SUPPORT	Other work supports current work or is supported by current work
USE	Other work is used in own work
FUT	Statements/suggestions about future work (own or general)

2.4　Multi-Layer Scientific Discourse 标注模型

该模型是 2015 年由法培拉大学自然语言处理团队的 B. Fisas、F. Ronzano 等结合计算机图形学领域实际情况，创新性地提出的简化版标注模型[27－28]。计算机图形学是一个相对年轻的学科，在语义标注上不同于生物科学有很成熟的词表标注，同时计算机图形学大多有技术背景，比如物理学、机械、流体动力学、数学等，因此模型更侧重于算法、方程式、代数和数学推理等。该模型也遵循了前人的研究，认为句子是表示语篇元素的最佳粒度，因此模型也是基于句子级别的标注。

为了加强对科技全文语篇元素以及整体科技文献的理解，研究人员综合考虑到语篇类别修辞、引用的价值、与文章中心思想接近程度、交叉特征标注等，共同形成了 Multi-Layer Scientific Discourse 多层次科技文献标注模型[28]。如图 3 所示，每一个句子都包含了 4 个层次的信息，从左向右依次是类别修饰层、引用目的层、交叉特征层、中心相关度层。

(1)语篇类别层次。主要包含了来源于对 CoreSC 模型和 AZ 模型的类别进行简化映射后最终确定的 5 个类别，分别是 Challenge、Background、Approach、Outcome、Future Work，具体类别定义和简化说明见图 4[28]。

(2)引用目的层次。这部分主要是对文献中的引用进行细化标注，主要采用了 A. Abu-Jbara 等[48]的标注模型提议，具体引用目的类别见表 3[28]，包括评论、对比、使用、基础工作或通用研究等方面，每一个类别拥有不同的子属性，如 Weakness 和 Strength 包含了评价极性，Evaluation 目的在于收集那些关于一篇引文正面和负面评论的句子；Similarity 和 Difference 是对比的对立原因；Use(引用)进行方法、数据或者工具引用标注区分等；Basis 类标注了作者引用自己的工作(Own Work)还是其他人的工作等。Neutral 类别包含了对研究者工作的描述、更多信息引用或者领域通用的实践等。

chinaXiv:202308.00603v1

ChinaXiv合作期刊

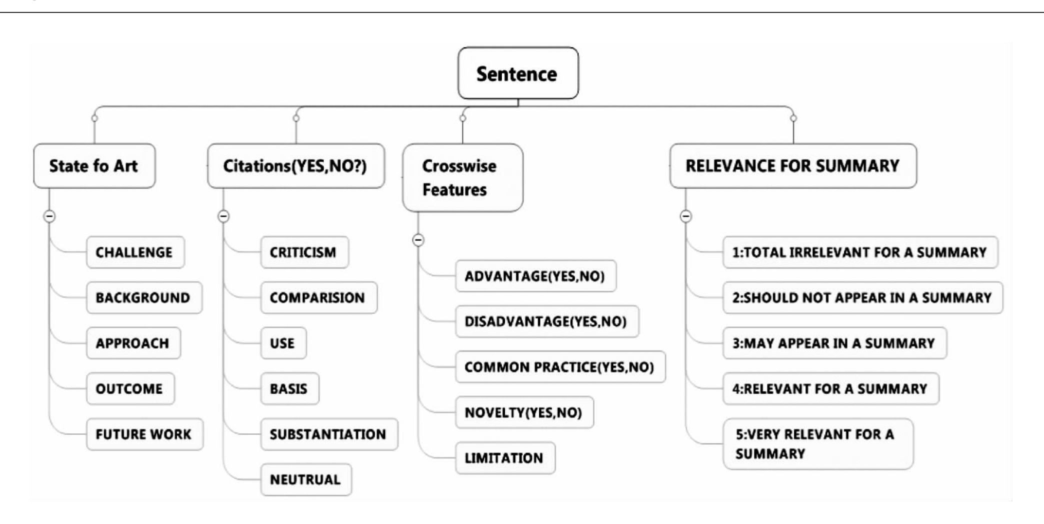


图 3　Multi-Layer Scientific Discourse 标注模型

CHALLENGE: The current situation faced by the researcher: it will normally include a Problem Statement, the Motivation, a Hypothesis and/or a Goal.
BACKGROUND:This section presents all the information which is helpful for understanding the situation or problem that is the subject of the publication. It will include sentences that state widely accepted knowledge in the domain (Common Ground) as well as previous related work (Related Work).
APPROACH: In this section the author explains HOW he intends to carry out the investigation. He may refer to a theoretical model or framework (Model), give some or many details of the experimental setup (Experiment), point to some data/phenomena observed during the experimentation (Observations) or comment on his decisions for choosing this methodology (Method).
OUTCOME: Here the author offers the study findings: measurable data without discussion (Results), an interpretation or analysis of the results in support of the conclusion (Discussion), how the research will contribute to the current knowledge in the field (Contribution) and an overall conclusion that should reject or support the research hypothesis (Conclusion). Any comments on the limitations of the authors work will also be included in the OUTCOME section.
FUTURE WORK: In most articles, the author will suggest or recommend further research to improve or extend his own work.

Figure 2: Description of the 5 categories of our Simplified Discourse Annotation Scheme

图 4　Multi-Layer Scientific Discourse 标注模型类别定义说明

表 3　引用目的类别

目的	子目的	目的	子目的
CRITISM	Weakness	SUBSTANTIATION	
	Strength	BASIS	Previous Own Work
	Evaluation		Others work
	Other		Future work
COMPARISON	Similarity	NEUTRAL	Description
	Difference		Ref. for more information
USE	Method		Common Practices
	Data		Other
	Tool		
	Other		

（3）交叉特征层。科技文献语篇特征 Advantage 和 Disadvantage 可以用来描述作者自己的方法和引用文献的特征，由于优劣势通常在一句话内出现，因此交叉特征层包含了双精度类别 Advantage-disadvantage 和 Disadvantage-advantage、创新（Novelties）和领域通用实践的特征标注。最后的 Limitations（局限性）特征仅指作者自己的工作，这在比较不同的调查研究上很重要。综上 5 个交叉特征类别为 ADVANTAGE、DISADVANTAGE、COMMON、NOVELTY、LIMITATION。

（4）中心相关重要度层次。按照每一个句子对中心思想贡献度设置了 5 个层级的分值，如完全不相关、不应该出现在文摘中、应该出现在文摘中、相关、非常相关，依次从 1 分到 5 分。

2.5　研究设计指纹描述模型

中国科学院文献情报中心钱力、张晓林等[34] 2014 年提出利用研究设计指纹对科技文献进行结构化描述，提升科技文献的计算机可识别性、可执行性，帮助科研人员快速了解科技文献的研究方法、算法、工具及结论等，并为未来的科学出版（即语义化出版）提供相应的出版规范参照。具体的研究设计指纹描述模型见表 4。研究设计指纹框架体系结构以研究设计指纹来表示科技文献研究成果，总体结构分为两个层次，第一个层次分为研究主题、研究方法、研究算法、研究结果、

chinaXiv:202308.00603v1

研究结论与未来研究六大部分；第二个层次详细描述科技文献，主要分为研究假说、研究场景、研究目的、研究背景、研究方法、研究数据、研究算法、研究结果、研究结论、未来研究以及研究设备共 11 种设计指纹，两个层次之间相互关联，层次内部相互关联，可很好地支持科技资源之间的关联计算与发行。该研究模型针对全文的 4 个粒度进行标注，即标题、摘要层，正文论证分区层，句子层，主题词层。

表 4　研究设计指纹描述模型

层次	子类型
研究设计层次	研究主题、研究方法、研究算法、研究结果、研究结论、未来研究
详细描述科技文献	研究假说、研究场景、研究目的、研究背景、研究方法、研究数据、研究算法、研究结果、研究结论、未来研究、研究设备

3　国外各个语篇元素标注模型和工作进展对比分析

由于上述模型研究中国外的模型都进行了系统建设和应用实现，因此本部分着重对国外的研究进展进行分析比较。首先从不同角度比较了 4 种模型构建的相同点和不同点，从而为对研究人员实际构建模型提供建议和参考，主要从模型的构建思想和任务、模型的类别以及构建方法、标注的领域和语料数据集合、标注工具和分类算法以及最终的实验效果分析几个角度去比较。最后也对模型研究中的问题进行了总结，以便为接下来的研究工作提供参考。

3.1　模型构建思想和解决任务的对比分析

模型构建思想作为研究者建立模型的初衷和支撑整个模型的理论研究基础非常关键，也决定了整个模型的区分粒度和划分角度。相同点就是每一个研究者都是为了从文献中更好地提炼挖掘语篇元素价值片段，不同点是研究者面临着不同的任务驱动，如 SciAnnoDoc 模型主要的目的是为了应用在检索系统中，提升检索效率，方便用户快速找到想要的知识片段；CoreSC 模型是为了更好地以本体研究的视角，全方面地解释一项调查研究工作；AZ 模型是基于知识声明观点，更加强调作者的贡献和引用他人的工作；Multilayer 模型则是顺应技术的发展，更好地解决新领域的文献语义分析问题，因此研究者要考虑解决实际的研究任务需要建立不同的研究模型，具体如表 5 所示。通过分析可发现，如果研究人员研究重心在于对一个学科领域的研究内容进行组织揭示，可采用 CoreSC 基于本体的揭示模型；如果研究人员侧重于去研究发现知识产权的相互影响和学者贡献影响，可采用 AZ 模型，方便区分他人与作者本人的贡献；研究人员如果面临实际对文献中心思想提炼、文献知识定向检索的应用场景，可采用 SciAnnoDoc 和 Multilayer 模型，以便帮助用户快速找到想要的知识片段。

表 5　模型构建思想和任务比较

模型名称	构建思想	解决任务
SciAnnoDoc	以用户为中心，回答用户精确或复杂的科学论文的查询问题，需要对文章的全部内容进行精确建模	提高精确检索效率
CoreSC	基于本体理论，认为科学论文是一项包含核心科学概念的科学调查研究内容的表示	揭示科技文献的丰富语义
AZ、AZ-II	基于知识声明观点，假设科技文献包含了作者、贡献者的知识产权和贡献说明	自动生成科技文摘，进行引文分析
Multilayer	综合 CoreSC 和 AZ 模型	自动分析科技文献语篇结构

3.2　语篇元素类别和建立方法对比分析

模型的建立方法大都经过不断的论证完善，如 SciAnnoDoc 先是通过调查问卷，然后再邀请专家确认最后类别；CoreSC 模型主要是基于对科学实体本体演化，选择核心概念类别，进而完成 CoreSC 类别确定；Multilayer 类别的确认在上述模型 1.4 中详细论述，也是精简了上述两个模型的 16 类类别和概念。模型标注粒度大部分是基于句子粒度的标注，如 CoreSC、AZ、Multilayer 模型，而 SciAnnoDoc 为了检索内容的丰富性选择了片段标注。具体见表 6。

本文进一步对模型的类别进行了详细对比，见图 5。各个模型不仅在类别上具有相似的或者相同的名字，如 BACKGROUND、AIM（GOAL）、METHOD（APPROACH），在具体标注的范围上也存在交叉覆盖映射。

3.2.1　CoreSC 模型与 AZ 模型对比　M. Liakata 和 S. Teufel 对 CoreSC 和 AZ 模型进行对比标注[11]，指出这两种模型在科技文献表示观点上形成互补，CoreSC 模型中的 BACKGROUND 包含了一般中立的背景知识同时也包含了现有知识声明，对应在 AZ_II 模型中分别为 OTHER、PREV_OWN 和 CO_GRO 类别。AIM 类在 AZ 模型中是研究目标的声明，但在 CoreSC 模型中可分解到以下 3 个类别中：GOAL（研究调查的目标状态）、HYPOTHESIS（尚未验证的声明）、OBJECT（研究

chinaXiv:201808.00003v1

ChinaXiv合作期刊

表 6 模型语篇元素类别和构建方法对比表

模型名称	类别数	粒度	语篇元素组成	建立方法
SciAnnotDoc	5	片段	FINDINGS、HYPOTHESIS、METHODOLOGY、RELATED WORK、DEFINITION	用户调研、实证研究
CoreSC	11	句子	HYPOTHESIS、MOTIVATION、BACKGROUND、GOAL、OBJECT、METHOD、EXPERIMENT、MODEL、OBSERVATION、RESULT、CONCLUSION	本体演化、专家论证
AZ	7	句子	OTHER、OWN、BACKGROUND、BASIS、CONTRAST、AIM、TEXTUAL	知识声明导向,修辞结构理论建立
AZ-II	15	句子	AIM、NOV_ADV、CO_GRO、OTHR、REV_OWN、OWN_MTHD、OWN_FAIL、OWN_RES、OWN_CONC、CODI、GAP_WEAKANTISUPP、SUPPORT、USE、FUT	基于 AZ 和实际工作扩展
Multilayer	5	句子	CHALLENGE、BACKGROUND、APPROACH、OUTCOME、FUTURE WORK	简化 CoreSC 和 AZ 类别

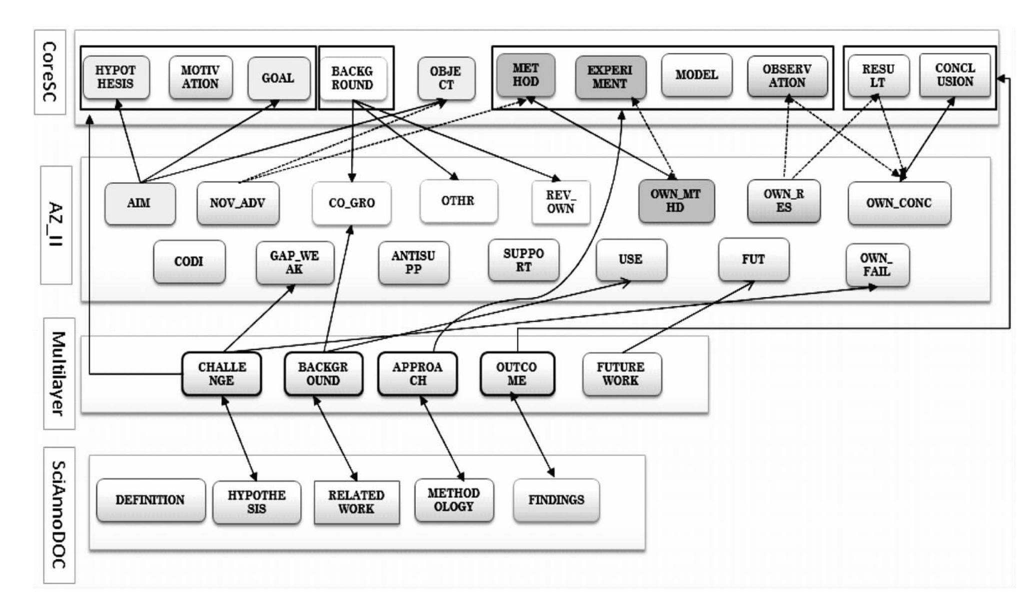


图 5 各个模型类别映射及对比

调查中相关联的特定实体或对研究调研实体规则属性的创新性或优劣势的声明)。OWN_MTHD 和 METHOD 都是指用到的方法,然而 CoreSC 进一步区分为实验方法(EXPERIMENT)、当前研究中使用的方法(Method-New)以及文章中提到的其他工作中使用的方法(Method-Old)。OWN_RES 与 CoreSC 类别中 OBSERVATION 相关,表示一项研究调查中的数据或现象记录。对比之下,CoreSC 类 RESULT 属于从 OBSERVATION 起源的事实论断。AZ-II 中 NOV_ADV 的类表示在文章中使用方法的新颖性和优势,对应到 CoreSC 中,可以对 METHOD 和 OBJECT 进行新颖性和优点的标注。其他的类别就是完全不同的类别,在 CoreSC 模型中 HYPOTHESIS、MOTIVATION、OBJECT 和 MODEL 是完全根据研究调查本体组织,而在 AZ-II 模型中 CODI、GAP_WEAK、SUPPORT、ANTISUPP、USE 和 FUT 则是按照与其他工作的关联组织,OWN_FAIL 阐述文献作者的不足之处。

3.2.2 CoreSC + AZ 模型与 Multilayer 模型对比 从 Multilayer 模型定义中可以很好地发现,该模型中的 CHALLENGE 类别表示当前研究者面临的研究状况,可以映射到 CoreSC 的 HYPOTHESIS、MOTIVATION 和 GOAL 以及 AZ 模型中当前没有解决的问题 GAP_WEAK 和 OWN_FAIL。BACKGROUND 揭示了对理解当前研究主体有用的发表信息,映射到 CoreSC 模型中 BACKGROUND 和 AZ 模型中的 CO_GRO 通用背景知识和使用的方法数据 USE 类别中。APPROACH 映射到 CoreSC 模型中可能包括了之前的 MODEL、EXPERIMENT、OBSERVATION、METHOD。OUTCOME 揭示了研究发现,包括可测量的数据结果(RESULT)或者结论(CONCLUSION)。FUTURE WORK 则对应 AZ 模型中的 FUT 类别。

3.2.3 Multilayer 模型与 SciAnnoDoc 模型对比 由于 SciAnnoDoc 模型完全是基于用户的角度去构建的,完全采用实证研究方法,因此模型更加接近人检索使用的角度。例如 DEFINITION 就是用户在希望了解一个领域知识的时候最基础的需求类别。其他的类别则基本与 Mutilayer 模型一一对应。

通过对比分析发现,自动标注模型的语义类别限

chinaXiv:202308.00603v1

定在6个就基本可满足各种工作需求,如研究目标、研究背景、研究方法、研究发现、研究结论以及包含科技文献中对关键术语的研究定义(概念解释),其他的各个类别可通过映射或转化到这些类别中去,这样可避免类别太多的冗杂和太少的宽泛问题,也可完全覆盖语义价值信息。

3.3 总体研究进展和成果对比分析

从各个模型研究领域和应用项目角度进行分析,体现了该项工作的具体研究价值和实用价值,也为接下来的研究者提供了丰富的语料和项目参考。具体如表7所示:

表7 总体研究项目和成果比较

模型名称	年份	应用(资助)项目	应用领域	项目成果
SciAnnoDoc	2011－今	瑞士自然科学基金	人文性别研究	A:1 400篇标注全文 B:FSAD应用系统
CISP/CoreSC	2007－2009	英国ART项目	生物化学	A:225篇标注全文 B:SAPIENT文章标注工具
	2010－2017	欧盟SAPIENT Automation项目		A:265篇标注全文 B:SAPIENTA自动化标注工具 C:应用与自动文摘系统 D:应用与CRA项目
AZ/AZ-II	2004－2007	剑桥大学研究性项目	计算语言学	A:80篇标注全文
	2010－今	CRA项目	生物化学领域	A:1 000篇标注摘要 B:CRAB在线阅读和标注系统
Multilayer	2016年	Dr. Inventor项目	计算机图形学	A:40篇标注全文 B:DRI Framework标注系统和框架

SciAnnoDoc研究团队致力于提升检索的效率,在瑞士自然科学基金项目的支持下[5-6],项目选择人文性别领域开发了一套面向用户的文献检索查询系统(FSAD系统),该系统基于提出的SciAnnoDoc模型对文献进行了标注,同时提供了相应的标注工具和人工标注语料,经过严格科学的用户使用评估证明,对比使用FSAD系统和基于传统关键字检索系统,使用前者用户解决问题的正确率和效率大大提升。

CoreSC模型的前身是CISP模型,研究始于2007年由英国高等教育联合信息服务委员会(JISC)资助的基于本体的文章表示工具项目(An ontology-based article preparation tool,简称ART),项目研究产出了一套基于CISP模型的手动标注工具SAPIENT,可方便快捷地基于决策树的方法实现人工标注,大大提升人工标注的效率和准度,也为构建可靠的训练语料奠定了坚实基础,项目实现了对225篇生物化学领域语料的人工标注[8-9]。为了实现机器自动化标注,2010年欧盟资助项目SAPIENT Automation继续上述研究,提出了CoreSC模型并基于该模型开发了自动标注工具SAPIENTA(注意比上边多了一个A,表示自动化Automation),语料也得到进一步丰富完善,完成265篇黄金标注数据集[11-12]。SAPIENTA具体应用于两个系统,一是自动文摘系统,实验证明CoreSC模型生成的文摘比Microsoft自动生成文摘效果更好,甚至在某些情况下要优于人工撰写的文摘;二是应用于生命科学研究领域CRA项目(癌症风险评估项目)[14-16]对该领域文章进行更好的标注,以方便研究人员对该领域论文深入研究分析。

AZ模型前期主要为剑桥大学的研究性项目,第一阶段由提出者联合相关人员对80篇计算语言学领域文章进行标注[20,22],并实现了机器自动标注。第二阶段应用到CRA项目,由剑桥大学教授A. Korhonen、瑞典卡洛琳斯卡医学院U. Stenius教授带领团队协助研究人员和风险评估人员的工作,有助于未来有效管理健康风险,项目使用AZ-II模型对相关文献进行自动标注,产出1 000篇摘要标注语料和CRAB 2.0标注工具,供癌症评估人员在线阅读和检索已经标注好的科研文献[15,38-39]。

Multilayer模型由欧盟委员会第七次框架项目资助,最终目的是利用科学技术的手段促进科技创新,侧重利用对科技文献的标注和知识挖掘发现可能的技术创新点。该项目最大的贡献在于完整地梳理和提出了同时集成很多开源工具的一套科技文献的标注框架和在线工具DRI Framework[26],方便相关研究人员参考使用,同时进一步应用在SKM Scientific Knowledge Miner项目进行科研知识挖掘工作[29-30]。

3.4 模型标注技术和分类效果的对比分析

经过上述分析,系统应用效果的好坏关键还依赖于语义类别标注的效果,本文比较了基于各个模型自动标注的实现效果,分别对正确率P、召回率R和F1值进行分析以及在自动标注过程中选用的分类方法[43]和最好、最差、平均值的效果比较。由于每个研究者的领域不同,因此实验数据集合也不同,本文提供每个研究者整体的实验效果,并不进行相同数据集的

chinaXiv:202308.00603v1

ChinaXiv合作期刊

实验分析比较(基于相同数据集的实验比较可作为进一步研究实施任务). 虽然基于不同的数据集,但本文从结果数据上进行整体分析,因此不影响本文即将进行的研究综述和结论分析。

具体而言,SciAnnoDoc 主要是依赖于人工撰写不同类别的语法规则[4],包括 20 个 Finding 规则、34 个 Definitions 规则、11 个 Hypothesis 规则和 19 个 Methodologies 规则,利用 1 400 篇人工标注进行训练,对 555 个句子实现了自动分类,分类效果见表 8。

表 8　SciAnnoDoc 分类结果

类别	句子个数	P	R	F1
Findings	168	0.82	0.39	0.53
Hypothesis	104	0.62	0.29	0.39
Definition	111	0.80	0.32	0.46
Methodologies	172	0.83	0.46	0.59
平均值	139	0.77	0.37	0.49

CoreSC 模型的研究人员分别利用不同的特征基于 265 篇语料进行了 10 – 交叉验证和实验,利用支持向量机 SVM、随机向量场 CRF 和线性核分类器进行自动分类实验,在此仅列举效果比较好的结果,利用所有特征值的基于 SVM 的分类器在各个类别的分类效果[12]如表 9 所示:

表 9　CoreSC 模型分类结果

类别	BAC	CON	EXP	GOA	MET	MOT	OBS	RES	MOD	OBJ	HYP	平均值
P	0.56	0.50	0.72	0.37	0.33	0.25	0.53	0.46	0.54	0.43	0.32	0.46
R	0.68	0.41	0.78	0.20	0.25	0.06	0.47	0.57	0.52	0.29	0.13	0.40
F1	0.62	0.45	0.75	0.26	0.29	0.10	0.50	0.51	0.53	0.34	0.19	0.41

由于 AZ-II 模型相对比较复杂,并且有一些类别是 AZ-II 所独有的,因此效果的好坏不具备与其他模型的可比性,因此本文选择了相对简单的 AZ 模型进行自动标注分类比较。实验基于 AZ 人工标注的 80 篇计算语言学文章,进行交叉实验验证,采用朴素贝叶斯分类器进行分类,取得的效果如表 10 所示[22]:

表 10　AZ 模型分类结果

类别	AIM	CONTR.	TEXTUAL	OWN	BACKG.	BASIS	OTHER	平均值
P	0.44	0.34	0.57	0.84	0.40	0.37	0.52	0.50
R	0.65	0.20	0.66	0.88	0.50	0.40	0.39	0.53
F1	0.52	0.26	0.61	0.86	0.45	0.38	0.44	0.50

Multilayer 模型在计算机图形学领域取得的分类效果见图 6,基于人工标注的 40 篇语料,分别选择了逻辑回归和 SVM 分类器进行测试,对相应的 F1 值进行了记录比较,结果逻辑回归取得的效果更好(缺少了正确率和召回率统计信息)。

综上,对各个模型整体对比情况如表 11 所示:

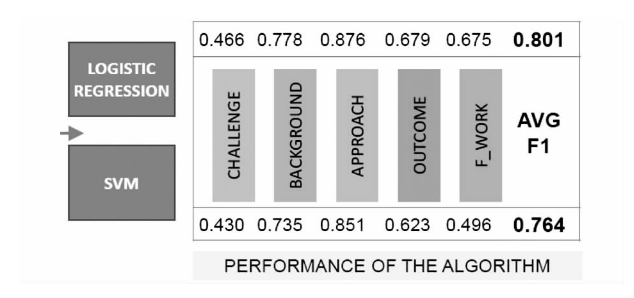


图 6　Multilayer 模型分类结果

表 11　模型标注工具和分类算法比较

模型名称	自动分类算法	最好效果(类别-P-R-F1)	最差效果(类别-P-R-F1)	平均效果(P-R-F1)
SciAnnoDoc	基于规则匹配算法	Methodology- 83% -46% -59%	Hypothesis-62% -29% -39%	77% -37% -49%
CoreSC	支持向量机 SVM、随机向量场 CRF	Expriment-72% -78% -75%	MOT-25% -6% -10%	46% -40% -41%
AZ/ AZ-II	综合 NB、SVM、CRF	OWN-81% -91% -85%	CONTR. -34% -20% -26%	50% -53% -50%
Multilayer	逻辑回归 LR 和 SVM	Approach-F 值——87.6%	CHALLENGE-F 值-47%	——F 值-80%

chinaXiv:202308.00603v1

3.5 语篇元素自动标注模型存在问题分析

(1)自动标注的语义类别种类以及蕴含意义定义问题难度较大,类别太少或太泛不能够满足用户的需要(如 SciAnnoDoc 和 Multilayer),但类别过多,不同的类别之间又会出现交叉覆盖,造成冗余和分类困难(如 CoreSC 和 AZ-II 模型),因此需要结合实际用户需要和研究需要进行反复的实证调查研究。由于每个人的知识背景和知识理解不同,调查研究充满了主观性和人为干扰因素,应制定尽可能适用于更多人群的类别,克服由于人的因素带来的不一致性存在难度。

(2)人工标注数据工作量大,耗时长,工作繁琐。每一个研究都需要人工进行初步的语料标注,并且对标注的精确度有较高要求,这一人工标注数据集的质量作为训练数据会直接影响到自动分类的效果。因此大多需要选用相应领域的专家进行语料人工标注,只有 AZ 模型为了降低人工标注数据的困难和准确性,建立了一个基于决策树的人工标注指导手册[20],方便人工标注,即使非本领域的专家也可以方便地按照决策树的指示完成标注,实现了非领域依赖,但大多数都还是领域依赖模型。

(3)最终自动分类的效果仍不是很理想,分类技术和方法仍有很大的提升空间。通过模型标注技术和分类效果分析发现无论是基于哪一个领域的语料数据和分类方法,当前研究自动标注模型的效果都不是特别理想,存在进一步的提升空间。SciAnnoDoc 模型与其他模型分类方法不同,主要是基于规则的方法,正确率相对较高,平均可达 75%,但召回率很低,平均只有 35%。为了提高召回率,需要写更多的规则,但是规则越多,噪音标注数据的风险越高。CoreSC 模型按照科学研究调查组织分类,对于描述核心概念比较完善,但是类别太多,有些类别之间容易混淆甚至本身就很模糊,会给人工标注数据带来困难,毕竟让专家对句子进行 11 个类别标注是个很难的工作,因此产生的训练数据本身的可靠性会是个问题,所得到的分类效果对比其他模型指标较低,F 值最高 75%,最低 10%,平均只有 41%。AZ 模型引入基于决策树的标注方法在一定程度上提高了训练数据的精准度,如在 OWN 类别分类上准确率和召回率分别可达 81% 和 91%,但 AZ 模型相对 AZ-II 模型本身比较落后,分类的角度也比较宽泛,在使用该模型过程中一般需要对模型进一步优化。Multilayer 模型研究领域和研究工作较新,基于前人的研究成果无论在模型上还是在分类方法上都明显优于其他模型,但在分类过程中选用了一些计算机图形领域特有的特征作为训练,所以对于其他领域的可扩展性有待进一步验证,并且类别较少,又有可能不足以满足研究人员的需要。

4 结论与展望

对科技文献进行语篇元素研究和标注工作在当前知识产权发展创新、知识出版业务创新、知识服务引领创新的时代背景下,有着重要的理论研究和实际应用价值,广泛应用于搜索引擎、自动文摘、科技创新点发现、自动问答系统[17]、语义出版[47]、写作教学[35]、网络语义知识组织、引文推荐[19]、日本法律条文标注[25]等知识服务环节。除此之外也广泛应用于医学、生命科学的科学研究工作中,如癌症风险评估、生命科学基因新功能对照/基因发现[45-46]、循证医学[23]等,实现多学科跨领域的研究合作。

本文详细研究比较了几个模型的研究工作,各团队针对不同的研究任务和侧重点,选择不同的领域进行了人工标注,产生了一系列标注数据集和方法集合,这些模型既互补又各有不同,这都为进行这方面领域研究提供了很有价值的参考。对科技文献语篇元素进行标注时,标注模型的确定是这项工作的基础部分,也是研究的核心和重要组成部分,首先需确定标注的任务和目标领域,不同的研究任务和领域由于研究内容本身和研究者思路本身的不同,文章的结构和内容也会千差万别,对模型的选择和类别区分也会明显不同。其次,选择标注的内容粒度,如基于片段、基于概念分区、基于句子、基于事件的标注。不同的粒度之间也不是孤立的,全文是由一个个片段组成,而一个个片段又是由句子组成,句子中又包含不同的事件。最后,确定模型和分类类别,一定要结合实际的应用情况灵活选择和制定。通常情况下,一个句子通常能较好地表达作者的意图,同时利用计算机技术手段可有效实现句子切分,也可有效避免基于片段的句子语义类别冲突的情况,因此可选择基于句子层级的标注进行研究。对于语义类别的个数,通过分析发现 5－7 个类别通常可包含所有的语义描述,本文结论为 6 个语义类别可覆盖大多数语篇元素语义类型,即研究目标、研究背景、研究方法、研究发现、研究结论以及研究定义。

科技语篇元素最终的自动化分类效果,直接决定

chinaXiv:201808.00060v1

ChinaXiv合作期刊

了应用的效果,是整个工作中的研究重点和技术难点部分。基于规则的分类可获得较高的准确率但召回率不太理想,基于机器学习的分类器需要大量的训练人工标注语料,对缺少人工支持的领域带来挑战,并且对有些类别的分类效果也不是很理想。当前人工智能领域的迅速发展使得机器学习和深度学习的方法进一步得到发展,已有研究人员开展弱监督[31,36]、无监督[40]、基于深度神经网络学习[42]的学习算法的研究以解决该类问题。也可创新融合各类算法进行分类,以进一步提高分类效果,这将作为本文作者接下来的重点研究工作。

本文对构建模型的基础理论假设和构建思想进行了总结对比,之前也有一些模型对比分析的研究,都缺少从这一角度进行比较,而理论基础或假设往往是开展一项研究的出发点,对后续研究和模型定型具有决定性作用。但该调研工作可能还存在不足和不够全面的地方,接下来笔者将进一步加强相应的调研工作。未来笔者会结合物理领域语义丰富化的工作具体建立适合该工作需要的模型并加以实现,同时可基于相同的数据集对上述不同的方法进行实验验证,观察相应的实验效果,使得对比工作更具参考价值,也希望提出创新性分类方法来解决上述分类问题。

chinaXiv:202308.00603v1

参考文献:

[1] FALQUET G. New trends for reading scientific documents[C]// ACM workshop on online books, complementary social media and crowdsourcing. New York: ACM, 2011:19–24.

[2] RIBAUPIERRE H, FALQUET G. A user-centric model to semantically annotate and retrieve scientific documents[C]//Proceedings of the sixth international workshop on exploiting semantic annotations in information retrieval. New York: ACM,2013: 21–24.

[3] RIBAUPIERRE H, FALQUET G. User-centric design and evaluation of a semantic annotation model for scientific documents[C]// Proceedings of the 14th international conference on knowledge technologies and data-driven business. New York: ACM,2014: 40.

[4] RIBAUPIERRE H. Precise information retrieval in semantic scientific digital libraries[D]// Genève:UNIVERSITÉ DE GENÈVE, 2014.

[5] RIBAUPIERRE H, FALQUET G. An automated annotation process for the SciDocAnnot scientific document model[C]//Proceedings of the 5th international workshop on semantic digital archives. Osaka: International Workshop on Semantic Digital Archives,2015:30–41.

[6] RIBAUPIERRE H, FALQUET G. Extracting discourse elements and annotating scientific documents using the SciAnnotDoc model: a use case in gender documents[J]. International journal on digital libraries, 2017,1(3):1–16.

[7] SOLDATOVA L N, KING R D. An ontology of scientific experiments[J]. Journal of the royal society interface, 2006, 3(11): 795–803.

[8] SOLDATOVA L, LIAKATA M. An ontology methodology and cisp-the proposed core information about scientific papers[EB/OL]. [2018–05–31]. http://repository. jisc. ac. uk/137/1/Report-CISP. pdf.

[9] LIAKATA M, SOLDATOVA L. Guidelines for the annotation of general scientific concepts[J]. Applied & environmental microbiology, 2008, 61(3):1020–1026.

[10] LIAKATA M, CLAIRE Q, SOLDATOVA L N. Semantic annotation of papers: interface & enrichment tool[C]//Proceedings of the BioNLP 2009 workshop. boulder. Colorado: Association for Computational Linguistics,2009: 193–200.

[11] LIAKATA M, TEUFEL S, SIDDHARTHAN A, et al. Corpora for the conceptualisation and zoning of scientific papers[C]// International conference on language resources and evaluation. Valletta: European Languages Resources Association (ELRA), 2010:105–108.

[12] LIAKATA M, SAHA S, DOBNIK S, et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications[J]. BMC bioinformatics, 2012, 28(7):991–1000.

[13] LIAKATA M, THOMPSON P, de WAARD A, et al. A three-way perspective on scientific discourse annotation for knowledge extraction[C]//Proceedings of the workshop on detecting structure in scholarly discourse. Jeju Island:Association for Computational Linguistics, 2012: 37–46.

[14] KORHONEN A, SILINS I, LIN S, et al. The first step in the development of text mining technology for cancer risk assessment: identifying and organizing scientific evidence in risk assessment literature[J]. BMC bioinformatics,2009, 10(1):1–19.

[15] GUO Y, KORHONEN A, LIAKATA M, et al. Identifying the information structure of scientific abstracts: an investigation of three different schemes[C]//Proceedings of the 2010 workshop on biomedical natural language processing. Uppsala:Association for Computational Linguistics, 2010: 99–107.

[16] GUO Y, KORHONEN A, LIAKATA M, et al. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment[J]. BMC bioinformatics, 2011, 12(1):1–18.

[17] LIAKATA M, DOBNIK S, SAHA S, et al. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. EMNLP. Seattle: Association for Computational Linguistics, 2013:

chinaXiv:202308.00603v1

747－757.

[18] RAVENSCROFT J, OELLRICH A, SAHA S, et al. Multi-label annotation in scientific articles-the multi-label cancer risk assessment corpus[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA), 2016.

[19] DUMA D, LIAKATA M, CLARE A, et al. Applying core scientific concepts to context-based citation recommendation[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA),2016.

[20] TEUFEL S, CARLETTA J, MOENS M. An annotation scheme for discourse-level argumentation in research articles[C]//Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics,1999: 110－117.

[21] TEUFEL S. Argumentative zoning: information extraction from scientific text[D]. Edinburgh: University of Edinburgh,1999.

[22] TEUFEL S, MOENS M. Summarizing scientific articles: experiments with relevance and rhetorical status[J]. Computational linguistics, 2002, 28(4): 409－445.

[23] TEUFEL S, BATCHELOR C, BATCHELOR C. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics[C]// Conference on empirical methods in natural language processing. Singapore: Association for Computational Linguistics,2009:1493－1502.

[24] HEFFERNAN K, TEUFEL S. Identifying problems and solutions in scientific text[J]. Scientometrics,2018(1):1－16.

[25] YAMADA H, TEUFEL S, TOKUNAGA T. Annotation of argument structure in Japanese legal documents[C]//Proceedings of the 4th workshop on argument mining. Copenhagen: Association for Computational Linguistics,2017: 22－31.

[26] RONZANO F, SAGGION H. Dr. inventor framework: extracting structured information from scientific publications[C]// JAPKOWICZ N, MATWIN S. Discovery science. Cham,: Springer, 2015: 209－220.

[27] FISAS B, RONZANO F, SAGGION H. On the discoursive structure of computer graphics research papers[C]//The 9th linguistic annotation workshop held in conjuncion with NAACL. Denver: Association for Computational Linguistics,2015: 42－51.

[28] FISAS B, RONZANO F, SAGGION H. A multi-layered annotated corpus of scientific papers[C]//Proceedings of the tenth international conference on language resources and evaluation. Paris:European Language Resources Association, 2016.

[29] RONZANO F, SAGGION H. Knowledge extraction and modeling from scientific publications[M]//Osborne: Springer International Publishing,2016:11－25.

[30] RONZANO F, FREIRE A, SAEZ-TRUMPER D, et al. Making sense of massive amounts of scientific publications: the scientific knowledge miner project[C]// BIRNDL 2016 joint workshop on bibliometric-enhanced information retrieval and NLP for digital libraries. New York: Digital Libraries. IEEE, 2016: 36－41.

[31] ANKE L E, SAGGION H, RONZANO F. Weakly supervised definition extraction[C]//Proceedings of the international conference recent advances in natural language processing. Shoumen: INCOMA Ltd, 2015: 176－185.

[32] 邢美凤. 科技文献中句子级新信息探测方法研究[D]. 北京: 中国科学院研究生院, 2012.

[33] 白光祖, 何远标, 马建霞, 等. 利用小样本量机器学习实现学术文摘结构的自动识别[J]. 现代图书情报技术, 2014, 30(7):34－40.

[34] 钱力, 张晓林, 王茜. 基于科技文献的研究设计指纹描述框架研究[J]. 大学图书馆学报, 2015 (1): 14－20.

[35] SONG W, FU R, LIU L, et al. Discourse element identification in student essays based on global and local cohesion[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon: Association for Computational Linguistics,2015: 2255－2261.

[36] GUO Y, KORHONEN A, POIBEAU T. A weakly-supervised approach to argumentative zoning of scientific documents[C]//Proceedings of the conference on empirical methods in natural language processing. Edinburgh:Association for Computational Linguistics, 2011: 273－283.

[37] CONTRACTOR D, GUO Y, KORHONEN A. Using argumentative zones for extractive summarization of scientific articles[C]// Proceedings of International Conference on Computational Linguistics. Mumbai:The COLING 2012 Organizing Committee, 2012: 663－678.

[38] SILINS I, KORHONEN A, GUO Y, et al. A text-mining approach for chemical risk assessment and cancer research[J]. Toxicology letters, 2014,229(4): S164－S165.

[39] GUO Y, SÉAGHDHA D O, SILINS I, et al. CRAB 2.0: a text mining tool for supporting literature review in chemical cancer risk assessment[C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. Dublin: Dublin City University and Association for Computational Linguistics,2014: 76－80.

[40] KIELA D, GUO Y, STENIUS U, et al. Unsupervised discovery of information structure in biomedical documents[J]. BMC bioinformatics, 2014, 31(7):1084－1092.

[41] BAKER S, SILINS I, GUO Y, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer [J]. BMC bioinformatics, 2015, 32(3): 432－440.

ChinaXiv合作期刊

[42] BAKER S, KORHONEN A. Initializing neural networks for hierarchical multi-label text classification[C]// 16th Workshop on Biomedical Natural Language Processing. Vancouver: Association for Comutational Linguistics, 2017:307 - 315.

[43] KIM S N, MARTINEZ D, CAVEDON L, et al. Automatic classification of sentences to support evidence based medicine[J] BMC bioinformatics, 2011, 12(2): S5.

[44] SOLLACI L B, Pereira M G. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey [J]. Journal of the medical library association, 2004, 92(3): 364 - 371.

[45] GOBEILL J, TBAHRITI I, EHRLER F, et al. Gene ontology density estimation and discourse analysis for automatic GeneRiF extraction[J]BMC bioinformatics, 2008, 9(3):S9 - 19.

[46] JIMENO-YEPES A J, STICCO J C, MORK J G, et al. GeneRIF indexing: sentence selection based on machine learning[J]. BMC bioinformatics, 2013, 14(1):1 - 10.

[47] CLARK T, CICCARESE P N, GOBLE C A. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications[J]. Journal of biomedical semantics, 2014, 5(1):28 - 61.

[48] ABU-JBARA A, RADEV D. Reference scope identification in citing sentences[C]//Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Montreal: Association for Computational Linguistics, 2012: 80 - 90.

作者贡献说明:

于改红:负责相关文献整理阅读和文章撰写;

张智雄:负责文章脉络把握和具体写作思路指导;

马娜:参与论文修改和撰写。

Overview of Science and Technology Literature Discourse Elements Automatic Annotation Model Research

Yu Gaihong[1,2] Zhang Zhixiong[1,2,3] Ma Na[1,2]

[1] University of Chinese academy of sciences, Beijing 100049

[2] National Science Library, Chinese Academy of Sciences, Beijing 100190

[3] Wuhan Library, Chinese Academy of Sciences, Wuhan 430071

Abstract: [**Purpose/significance**] In order to improve the semantic enrichment effect of scientific and technical literature, this paper summarizes the domestic and foreign scientific and technical literature discourse elements automatic annotation model, technologies and methods, and provides reference for text mining, knowledge extraction and semantic analysis system. [**Method/process**] This paper used Web Scholar and related database search engine to conduct in-depth reading and related research on references and research reports involving scientific and technical papers annotation, discourse elements, knowledge extraction, sentence recognition, automatic article classification, etc. and summarized the research the main technologies of each module in the framework. [**Result/conclusion**] The annotation of scientific literature discourse elements has very important practical application value. The construction of annotation model needs to take full account of construction thought, annotation field and annotation granularity as well as annotation techniques.

Keywords: scientific and technical literature discourse elements annotation model automatic annotation

chinaXiv:202308.00603v1